From:	POULSEN Mike
To:	Eric Blischke/R10/USEPA/US@EPA
Cc:	Karl Gustavson/DC/USEPA/US@EPA; Burt Shephard/R10/USEPA/US@EPA; Chip Humphrey/R10/USEPA/US@EPA; GAINER Tom; Gina Grepo-Grove/R10/USEPA/US@EPA; PETERSON Jenn L; jeremy_buck@fws.gov; ANDERSON Jim M; Joe Goulet/R10/USEPA/US@EPA; Judy Smith/R10/USEPA/US@EPA; Kristine Koch/R10/USEPA/US@EPA; MCCLINCY Matt; Rene Fuentes/R10/USEPA/US@EPA; Robert.Neely@noaa.gov; Sean Sheldrake/R10/USEPA/US@EPA; tomd@ctsi.nsn.us; rose@yakama.com; erin.madden@gmail.com; jay.field@noaa.gov; Lori Cora/R10/USEPA/US@EPA; Mark Ader/R10/USEPA/US@EPA; audiehuber@ctuir.com; Lisa.Bluelake@grandronde.org; sheila@ridolfi.com; Benjamin Shorr; LavelleJM@cdm.com; Mary.Baker@noaa.gov; Michael.Karnosh@grandronde.org; FARRER David.G; dallen@stratusconsulting.com; jpeers@stratusconsulting.com; Bob Dexter; cunninghame@gorge.net; JMalek@parametrix.com; nancy.munn@noaa.gov; jweis@hk-law.com; Brad Hermanson; frenchrd@cdm.com; ryan@davissudbury.com; Judy Smith/R10/USEPA/US@EPA; Genevieve.Angle@noaa.gov; TARNOW Karen E; Jessica.Winter@noaa.gov; mspence@parametrix.com; Elizabeth Allen/R10/USEPA/US@EPA
Subject:	Portland Harbor - SQG Comparison
Date:	03/29/2010 04:24 PM
Attachments:	PH FPM SQG comparison.xlsx

Eric -

As discussed in the TCT last week, I am sending a spreadsheet showing a comparison of high-screen SQGs. High-screen SQGs are levels above which we expect to see benthic toxicity, as distinct from low-screen SQGs, below which we expect to see no toxicity. One set of high-screen SQGs was developed by the LWG using their approach to the floating percentile method, and the other set of SQGs was derived by us following EPA's 2009 direction to the LWG. Specifically, the EPA approach uses different bioassay hit criteria, and involves looking at results from all four bioassay endpoints (including Hyalella biomass), and taking the lowest values. The EPA values were obtained from spreadsheets provided by the LWG. I encourage people to check the values.

As you can see from the backup worksheets, the different chemical lists for the endpoints had to be adjusted to calculate reliabilities. Reliability was evaluated using the pooled dataset (a hit in any bioassay constitutes a hit for the sample).

Sometimes SQGs developed using the EPA approach were lower than those developed using the LWG approach, sometimes they were the same, and occasionally they were higher. DDX is an example of the EPA SQG being much higher than the LWG SQG. This is possibly an artifact of the SQG for DDX being developed in conjunction with DDD, DDE, and DDT. If, for example, DDT and DDX co-vary as we expect they would, it is possible to have the SQG for one low and the SQG for the other chemical higher, or the other way around. The model would not make a distinction about which set of SQGs is "correct" if the reliability goals are met. We may want to include SQGs for DDD, DDE, DDT, and total DDX to account for covariance.

The DDT result (is 72 ug/kg or 8110 ug/kg the correct SQG?) illustrates one of the concerns Jennifer Peterson and I have about developing SQGs using the FPM. Covariance may also be an issue with PAHs or other constituents. In addition, we suggest that the LWG look at sources of variability, such as bioavailability. Jay found that organic carbon normalization is important for improving relationships using the logistic regression method for developing sediment SQGs. Also, when we have the requested appropriate input files (and the time), we will look at how much the number of chemicals included, which chemicals, and their order will impact the development of SQGs using the FPM.

In addition, given the dataset used to derive both low and high screening values, we found that we are unlikely to meet the false negative and false positive goals of 20%, and also reach the predicted hit reliability goal of 80% for high values. This is because the dataset contains mostly no-hit results. Even with a relatively good false positive rate, because there are lots of no-hit samples, there will be lots of predicted-no-hit samples relative to the number of predicted-hit samples. Early in the process we knew that predicted-hit and predicted-no-hit measures depend on the mix of hit and no-hit results.

There is a related issue that could affect reliability measurements. In using a high screen to identify areas requiring remediation, the samples are more likely to be contaminated than would be expected from the general database. The actual predicted-hit reliability rate, when calculated using the more contaminated dataset, should be greater than the predicted-hit reliability calculated using the SQG spreadsheet. In fact, false negative and false positive rates could also change when a more contaminated dataset is used. The false negative rate as calculated in the SQG spreadsheet could be overestimated at the high level (and underestimated at the low level). I have a figure at the end of the spreadsheet that may help illustrate how this could occur.

Until these issues are resolved, Jennifer Peterson and I conclude that it is difficult to evaluate reliability of any set of SQGs with the current dataset used to develop the SQGs. Our recommendation is to use an appropriate validation dataset to evaluate reliability. This was a specific recommendation by Don MacDonald in his 2008 review of methods for developing Portland Harbor SQGs. Without knowing the reliability of SQGs based on a validation dataset, and until our questions on the FPM approach are resolved, we are reluctant to use site-specific benthic toxicity SQGs as PRGs. But if site-specific SQGs are going to be used, we suggest that we use the SQGs developed from the EPA approach and not

the LWG approach.

If you see any errors in the tables, or if you have questions, please let me know.

- Mike